IRIM at TRECVID 2010: Semantic Indexing and Instance Search

David Gorisse, Frédéric Precioso, Philippe-Henri Gosselin, Lionel Granjon, Denis Pellerin, Michèle Rombaut, Hervé Bredin, Lionel Koenig, Rémi Vieux, Boris Mansencal, et al.

IRIM at TRECVID 2010: Semantic Indexing and Instance Search

David Gorisse[1], Frédéric Precioso[1,8], Philippe Gosselin[1], Lionel Granjon[2], Denis Pellerin[2], Michèle Rombaut[2], Hervé Bredin[3], Lionel Koenig[3], Rémi Vieux[4], Boris Mansencal[4], Jenny Benois-Pineau[4], Hugo Boujut[4], Claire Morand[4], Hervé Jégou[5], Stéphane Ayache[6], Bahjat Safadi[7], Yubing Tong[7], Franck Thollard[7], Georges Quénot[7], Matthieu Cord[8], Alexandre Benoît[9], and Patrick Lambert[9]

[1]ETIS UMR 8051, ENSEA / Université Cergy-Pontoise / CNRS, Cergy-Pontoise Cedex, F-95014 France
[2]GIPSA-lab UMR 5216, CNRS / Grenoble INP / UJF-Grenoble 1 / U. Stendhal-Grenoble 3 / 38402 Grenoble, France
[3]IRIT / UMR 5505 / Universit Paul Sabatier / F-31062 Toulouse CEDEX 9
[4]LABRI UMR 5800, Université Bordeaux 1 / Université Bordeaux 2 / CNRS / ENSEIRB, Talence Cedex, France
[5]INRIA Rennes / IRISA UMR 6074 / TEXMEX project-team / 35042 Rennes Cedex
[6]LIF UMR 6166, CNRS / Université de la Méditerranée / Université de Provence, F-13288 Marseille Cedex 9, France
[7]UJF-Grenoble 1 / UPMF-Grenoble 2 / Grenoble INP / CNRS, LIG UMR 5217, Grenoble, F-38041, France
[8]LIP6 UMR 7606, UPMC - Sorbonne Universités / CNRS, Paris, F-75005 France
[9]LISTIC, Domaine Universitaire, BP 80439, 74944 Annecy le vieux Cedex, Franc

Abstract

The IRIM group is a consortium of French teams working on Multimedia Indexing and Retrieval. This paper describes our participation to the TRECVID 2010 semantic indexing and instance search tasks. For the semantic indexing task, we evaluated a number of different descriptors and tried different fusion strategies, in particular hierarchical fusion. The best IRIM run has a Mean Inferred Average Precision of 0.0442, which is above the task median performance. We found that fusion of the classification scores from different classifier types improves the performance and that even with a quite low individual performance, audio descriptors can help. For the instance search task, we used only one of the example images in our queries. The rank is nearly in the middle of the list of participants. The experiment showed that HSV features outperform the concatenation of HSV and Edge histograms or the Wavelet features.

1 Semantic Indexing

1.1 Introduction

The classical approach for concept classification in images or video shots is based on a three-stage pipeline: descriptors extraction, classification and fusion. In the first stage, descriptors are extracted from the raw data (video, image or audio signal). Descriptors can be extracted in different ways and from different modalities. In the second stage, a classification score is generated from each descriptor and, for each image or shot, and for each concept. In the third stage, a fusion of the classification scores obtained from the different descriptors is performed in order to produce a global score for each image or shot and for each concept. This score is generally used for producing a ranked list of images or shots that are the most likely to contain a target concept.

1.2 Evaluation of image descriptors

Nine IRIM participants (ETIS, GIPSA, IRIT, LABRI, LEAR, LIF, LIG, LIP6 and LISTIC) provided descriptors and two participants (LIF and LIG) provided classification results using them allowing for comparing the relative performances of these descriptors. These descriptors do not cover all types and variants but they include a significant number of different approaches including state of the art ones and more exploratory ones.

We have evaluated a number of image descriptors for the indexing of the 130 TRECVID 2010 concepts. This was done within the development set that was split into two parts, one for training and the other for evaluation ("1-fold cross-validation"). We used the annotations provided by the TRECVID 2010 collaborative annotation organized by LIG and LIF [1]. The performance is measured by the Mean Average Precision (MAP) computed on the 130 concepts. Three types of classifiers were used for the evaluation:

"standard SVM" (LIF_SVM), "multi-learner SVM" (LIG_MSVM) and kNN with two variants depending upon whether the hyper-parameters avec tuned independently for each concept (LIG_KNNC) or globally for all concepts (LIG_KNNG).

We evaluated the following image descriptors:

- **ETIS/global_<attr><histogram_type><dict size>**: histograms computed for different visual attributes and dictionary size.
  <attr> = lab: LAB colors, qw: norm of quaternionic wavelets coefficients, 3 scales.
  <type hist> = : m1 × 1: histogram computed on the whole image, m1 × 3: 3 histograms on 3 vertical stripes, m2 × 2: 4 histograms on four image quarters[2, 3].
- **LIG/h3d64** : normalized RGB Histogram 4 × 4 × 4 (64-dim).
- **LIG/gab40** : normalized Gabor transform, 8 orientations × 5 scales (40-dim).
- **LIG/hg104** : early fusion (concatenation) of LIG/h3d64 and LIG/gab40 (104-dim).
- **LIG/opp_sift_har** : bag of word, opponent sift with Harris-Laplace detector [4], generated using Koen Van de Sande's software (4000-dim).
- **LIG/opp_sift_dense** : bag of word, opponent sift with dense sampling [4], generated using Koen Van de Sande's software (1000-dim).
- **IRIT/MFCC-average**: Mean of MFCC on homogeneous segments on the shot → 12 dimensions,
- **LABRI/residualMotion_NPI:** mean absolute residual motion vectors (x and y coords) for image divided in 8 × 8 blocks, normalized on each key frame → 128 dimensions.
- **LABRI/residualMotion_NPM:** mean absolute residual motion vectors (x and y coords) as in LABRI/residualMotion_NPI but normalized on the whole video.
- **LABRI/faces:** OpenCV+median temporal filtering, ratio of overlapping between block and face bounding box on image divided into 8×8 blocks → 64 dimensions
- **LISTIC_Stip1:** for each keyframe, the number of Spatio-Temporal Interest Points is provided.
- **LISTIC_Stip89:** acknowledge for the number of Spatio-Temporal Interest Points (STIPS) for 89 frames in the neighborhood of a key frame. The neighborhood is centered on a key frame and its size is 100, but the first 11 frames are removed due to the initialization of the STIPS computation.
- **LEAR_sift_bow4096**: Bag Of SIFT Words vectors with dict_size = 4096.
- **LIF_percepts_<X>_<Y>_1_15**: Intermediate level descriptor contains the prediction scores of 15 visual concepts a X × Y grid (X.Y.15-dim).
- **GIPSA_AudioSpectro_b28**: spectral profile in 28 bands on a Mel scale.
- **GIPSA_AudioSpectroN_b28**: spectral profile in 28 bands on a Mel scale, normalized.

Table 1 shows the relative performance of a number of descriptor combinations. Size is the number of dimensions of the descriptor vector. Not all combinations were used but in the cases where the comparison is possible and with a few exceptions, all methods have comparable performances though with significant variations.

1.3 Performance improvement by fusion of descriptor variants and classifier variants

In a previous work, LIG introduced and evaluated the fusion of descriptor variants for improving the performance of concept classification. We previously tested it in the case of color histograms in which we could change the number of bins, the color space used, and the fussiness of bin boundaries. We found that each of these parameters had an optimal value when the others are fixed and that there is also an optimal combination of them which correspond to the best classification that can be reached by a given classifier (kNN was used here) using a single descriptor of this type. We also tried late fusion of several variants of non-optimal such descriptors and found that most combinations of non-optimal descriptors have a performance which is consistently better than the individual performance of the best descriptor alone. This was the case even with a very simple fusion strategy like taking the average of the probability scores. This was also the case for hierarchical late fusion. In the considered case, this was true when fusing consecutively according to the number of bins, to the color space and to the bin fuzziness. Moreover, this was true even if some variant performed less well than others. This is particularly interesting because descriptor fusion is known to work well when descriptors capture different aspects of multimedia content (e.g. color and texture) but, here, an improvement is obtained using many variants of a single descriptor. That may be partly due to the fact that the combination of many variant reduces the noise. The gain is less

Table 1: Performance of the classifier and descriptor combinations

Descriptor	size	LIF_SVM	LIG_KNNC	LIG_KNNG	LIG_MSVM
ETIS/global_labm1x1x64	64		0.0380	0.0395	
ETIS/global_labm1x1x128	128		0.0390	0.0421	0.0268
ETIS/global_labm1x1x192	192		0.0402	0.0415	
ETIS/global_labm1x1x256	256		0.0402	0.0415	
ETIS/global_labm1x3x64	192		0.0488	0.0499	
ETIS/global_labm1x3x128	384		0.0501	0.0501	0.0413
ETIS/global_labm1x3x192	576		0.0496	0.0496	
ETIS/global_labm1x3x256	768	0.0295	0.0505	0.0504	
ETIS/global_labm2x2x64	256		0.0496	0.0491	
ETIS/global_labm2x2x128	512		0.0490	0.0501	
ETIS/global_labm2x2x192	768		0.0502	0.0490	
ETIS/global_labm2x2x256	1024	0.0252	0.0504	0.0491	
ETIS/global_qwm1x1x64	64		0.0398	0.0411	
ETIS/global_qwm1x1x128	128		0.0408	0.0428	0.0382
ETIS/global_qwm1x1x192	192		0.0415	0.0443	
ETIS/global_qwm1x1x256	256	0.0354	0.0415	0.0443	
ETIS/global_qwm1x3x64	192		0.0486	0.0494	
ETIS/global_qwm1x3x128	384		0.0511	0.0506	0.0554
ETIS/global_qwm1x3x192	576		0.0510	0.0510	
ETIS/global_qwm1x3x256	768		0.0512	0.0513	
ETIS/global_qwm2x2x64	256		0.0497	0.0504	
ETIS/global_qwm2x2x128	512		0.0529	0.0519	
ETIS/global_qwm2x2x192	768		0.0530	0.0528	
ETIS/global_qwm2x2x256	1024		0.0546	0.0531	
LIG/h3d64	64		0.0368	0.0367	
LIG/gab40	40		0.0302	0.0311	
LIG/hg104	104	0.0348	0.0524	0.0506	0.0534
LIG/opp_sift_har	4000	0.0422	0.0474	0.0490	0.0601
LIG/opp_sift_dense	1000		0.0562	0.0544	0.0544
IRIT/mfcc_average	13	0.0041	0.0175	0.0178	0.0022
LaBRI/residualMotion_nPI	128		0.0020	0.0020	
LaBRI/residualMotion_nPM	128		0.0020	0.0018	
LaBRI/faces	64		0.0022	0.0021	
LISTIC/Stip_1	1		0.0028	0.0023	0.0016
LISTIC/Stip_89	89		0.0113	0.0110	0.0049
LEAR/sift_bof4096	4096	0.0797	0.0678	0.0715	
LIF/pammcol_20_13_32_9	2340		0.0326	0.0327	
LIF/percepts_20_13_1_15	3900	0.0487	0.0506	0.0496	
LIF/percepts_10_6_1_15	900	0.0551	0.0536	0.0535	
LIF/percepts_5_3_1_15	225	0.0437	0.0528	0.0527	
LIF/percepts_2_2_1_15	60	0.0366	0.0453	0.0458	
LIF/percepts_1_1_1_15	15	0.0212	0.0352	0.0357	
GIPSA/AudioSpectro_b28	28		0.0188	0.0190	0.0062
GIPSA/AudioSpectroN_b28	28		0.0214	0.0226	0.0123

than when different descriptor types are used but it is still significant.

We have then generalized the use of the fusion of descriptor variants and we evaluated it on other descriptors and on TRECVID 2010. We made the evaluation on descriptors produced by the ETIS partner of the IRIM group. ETIS has provided 3 $\times$ 4 variants of two different descriptors (see the previous section). Both these descriptors are histogram-based. They are computed with four different number of bins: 64, 128, 192 and 256; and with three image decomposition: 1x1 (full image), 1x3 (three vertical stripes) and 2x2 (2 by 2 blocks). Hierarchical fusion is done according to three

levels: number of bins, image decomposition and descriptor type.

Table 2 shows the result obtained for fusion within a same descriptor type (fusion levels 1 and 2) and between descriptor types (fusion level 3). The fusion of the descriptor variants varies from about 5 to 10% for the first level and is of about 4% for the second level. The gain for the second level is relative to the best result for the first level so both gain are cumulated. For the third level, the gain is much higher as this could be expected because, in this case, we fuse results from different information sources. The gain at level 3 is also cumulated with the gain at the lower levels.

In the spirit of gaining the maximum from each information source (from each descriptor type here), we also tried to fuse the output of different classifier types (or variants) for a single descriptor or for combinations of descriptors or of descriptor variants. While the previous experiment was done using only one type of classifier (kNN), we now use additionally outputs from other classifiers (SVM and MSVM). As can be seen in table 1, only a limited number of classification results have been computed (SVM and MSVM are significantly slower).

Table 2 shows the result obtained when fusing additionally the output from other classifiers. Even though their output was available for only a small number of descriptor variants were available and the performance of SVM classifiers are often lower, the fusion almost always leads to a significant improvement. An improvement occurs at all level of descriptor variant fusion and, again, the gain is cumulated. At the last level, the gain is still of about 10%.

Both descriptor variants fusion and classifier variants fusion yields a significant improvement and these improvements cumulates. However, this method has a drawback: the volume of computations involved increases in a multiplicative way according to the number of descriptor variants and to the number of classifier variants. In the case of descriptor variants, there may be a multiplicative factor for each dimension according to which the descriptor may by varied. This multiplicative factor is applied to a value which is already high considering that the classifiers have to be trained and to be used for prediction for a large number of images or video shots and for a large number of concepts. Future work will be needed to investigate whether a similar gain can be obtained by using only a limited number of combinations and on how to choose them.

1.4 Hierarchical fusion

Hierarchical fusion with multiple descriptor variants and multiple classifier variants was used and optimized for the semantic indexing task. We made several experiment in order to evaluate the effect of a number of factors. We optimize directly the first levels of the hierarchical fusion using uniform or average-precision weighting. The fusion was made successively on variant of the same descriptors, on variant of classifiers on results from the same descriptors, on different descriptors and finally on the selection of groups of descriptors.

Table 4 describes the combination used for the IRIM runs submitted at TRECVID 2010. These runs are classified according to the expected ranking of systems. IRIM-4 and IRIM-2 uses visual descriptor only excluding face-based ones). Then audio descriptors are added for IRIM-3 and IRIM-1. For each of these pairs of runs, the difference between the first and the second is the method used of the final fusion stage: for the first one an average precision weighting is used while for the second, weights are determined by direct optimization by cross-validation. The second may lead to better performance but is more prone to over-fit to the data.

Table 5 shows the results of the parameter optimization and the performance of the different combination on the development set. All steps lead to an improvement on the development data, either horizontally or vertically (in the above table). As can be seen in the following sections, this is also generally the case on the test data.

1.4.1 Performances on the semantic indexing task

Table 6 presents the result obtained by the four runs submitted. Although the absolute performances are quite different between the one obtained during the cross validation step, the ranking of the run is almost the same. We found a bug in our submission: three pairs of concepts were swapped (due to a change in the alphabetical order). The first part of the result (submitted) indicates our performance with this bug (the official one). The second part (corrected) indicates our performance when the bug is corrected.

2 Instance Search

2.1 Task presentation

Instance Search (INS) is a pilot task introduced by NIST in TRECVid 2010 Campaign. Given visual examples of entities, it consists in finding segments of videos in the data set which contain theses instances. This year, there were 22 entities spread in 4 types: person, character, object, location. They are listed in table 7. IRIM participated in INS task definition by proposing two instances: tank (9021) and Willem Wever van (9022). As this year the intention was only to explore task definition and evaluation, only a rough estimate of searched instances locations was asked. Indeed, we had only to find the shots were the instance appeared,

Table 2: Fusion of descriptor variants. Best desc: performance of the best individual descriptor variant. Desc. fusion: performance of the fusion of variants.

Fusion output	Fusion level	Best desc.	Desc. fusion	Gain
ETIS/global_labm1x1x_all	1	0.0421	0.0447	+6.2%
ETIS/global_labm1x3x_all	1	0.0505	0.0563	+11.5%
ETIS/global_labm2x2x_all	1	0.0504	0.0559	+10.9%
ETIS/global_lab_all_all	2	0.0563	0.0584	+3.7%
ETIS/global_qwm1x1x_all	1	0.0443	0.0460	+3.8%
ETIS/global_qwm1x3x_all	1	0.0513	0.0550	+7.2%
ETIS/global_qwm2x2x_all	1	0.0546	0.0584	+7.0%
ETIS/global_qw_all_all	2	0.0584	0.0609	+4.3%
ETIS/global_all_all_all	3	0.0609	0.0801	+31.5%

Table 3: Fusion of classifier variants. Best DVF: performance of the best descriptor variant fusion (from the previous table). Class. fusion: performance of the fusion of classifiers. (1): no additional classifier was available for any descriptor variant. (2): only one additional classifier was available for only one descriptor variant.

Fusion output	Fusion level	Best DVF	Class. fusion	Gain
ETIS/global_labm1x1x_all	1	0.0447	0.0494	+10.5%
ETIS/global_labm1x3x_all	1	0.0563	0.0628	+11.5%
ETIS/global_labm2x2x_all	1	0.0559	0.0594	+6.3%
ETIS/global_lab_all_all	2	0.0584	0.0646	+10.6%
ETIS/global_qwm1x1x_all	1	0.0460	0.0531	+15.4%
ETIS/global_qwm1x3x_all	1	0.0550	0.0550	0% (1)
ETIS/global_qwm2x2x_all	1	0.0584	0.0584	0% (2)
ETIS/global_qw_all_all	2	0.0609	0.0645	+5.9%
ETIS/global_all_all_all	3	0.0801	0.0878	+9.6%

Table 4: Runs submitted at TRECVID 2010

Weighting method	AP	Opt.
Visual only	IRIM-4	IRIM-2
Same plus audio	IRIM-3	IRIM-1

Table 5: One-fold cross-validation result of the fusion process

Run	MAP	last fusion level method
ALL_visual_map (IRIM-4)	0.1304	Average Precision weighting
ALL_irim_map (IRIM-3)	0.1339	Average Precision weighting
ALL_visual_opt (IRIM-2)	0.1360	Direct optimization weighting
ALL_irim_opt (IRIM-1)	0.1403	Direct optimization weighting

but not the precise frame or the precise location of the instance in the frame.

2.2 Search method

We used a Region Based Similarity Search based on the original idea of Bag of Visual Words indexing of Sivic and Zisserman[5]. In the original work, images from videos are indexed using a codebook of visual words on the basis of local interest point (SIFT like) description of image content. Traditional retrieval method inspired by the Bag Of Words representation in the Text Retrieval community can hence be applied to their visual counterparts in the image. The Region Based Similarity Search aims to compute a codebook of visual words not relying on the SIFT or SIFT like interest point de-

Table 6: InfMAP result and rank on the test set for all the 30 TRECVID 2010 concepts (full task)

System/run	MAP submitted	MAP corrected	rank submitted	rank corrected
F_A_IRIM_RUN01_1	0.0442	0.0471	42	37
F_A_IRIM_RUN03_3	0.0434	0.0466	45	39
F_A_IRIM_RUN02_2	0.0415	0.0444	47	44
F_A_IRIM_RUN04_4	0.041O	0.0443	49	45

Table 7: Instances for TRECVid 2010

number	type	text	number of examples
9001	PERSON	George W. Bush	5
9002	PERSON	George H. W. Bush	4
9003	PERSON	J. P. Balkenende	5
9004	PERSON	Bart Bosh	5
9005	CHARACTER	Professor Fetze Alsvanouds from the University of Harderwijk (Aart Staartjes)	5
9006	PERSON	Prince Bernhard	5
9007	CHARACTER	The Cook (Alberdinck Thijn: Gijs de Lange)	5
9008	PERSON	Jeroen Kramer	5
9009	CHARACTER	Two old ladies, Ta en To	5
9010	CHARACTER	one of two officeworkers (Kwelder of Benema en Kwelder: Harry van Rijthoven)	5
9011	PERSON	Colin Powell	3
9012	PERSON	Midas Dekkers	5
9013	OBJECT	IKEA logo on clothing	5
9014	CHARACTER	Boy Zonderman (actor in leopard tights and mesh top: Frank Groothof)	4
9015	OBJECT	black robes with white bibs worn by Dutch judges and lawyers	3
9016	OBJECT	zebra stripes on pedestrian crossing	4
9017	OBJECT	KLM Logo	2
9018	LOCATION	interior of Dutch parliament	4
9019	OBJECT	Kappa Logo	5
9020	OBJECT	Umbro Logo	5
9021	OBJECT	tank	3
9022	OBJECT	Willem Wever van	5

scriptors, but on image regions resulting from image segmentation. In this work, we proposed the elementary segmentation of image into rectangular cells forming a grid. The grid parameters such as the number of cells $N = n * k$, where n and k are the number of cells per line and per column, defines the size of cells as W/n and H/k. These W and H stand for image width and height respectively. For spatial resolution of Sound and Vision data set used in INS task (352x288), the size of a cell was 35x28 pixels. An example of an instance of a concept 9001 is presented in figure 1a with the cell grid superimposed. Hence, a cell is considered as a semi-local picture element as its size allows for meaningful computation of a global statistical feature. The similar approach by image blocks or Numceps [6] proved to be rather efficient for retrieval task in former TRECVid competitions.

2.2.1 Statistical Global Features

In this work, we considered three global features for each cell:

- HSV histogram, quantized to 45+32+32 bins
- Wavelet histogram (YUV space), with adaptive quantizing

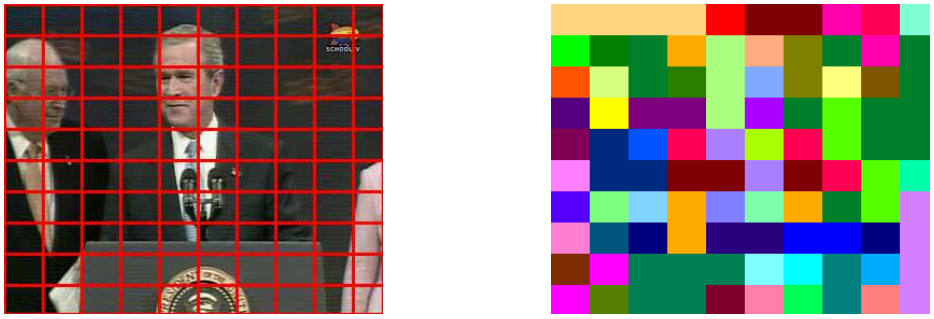

Figure 1: Semi-local picture elements (a) and assorted BOF (b)

- MPEG-7 Edge histogram

All these three features are histograms. Except MPEG-7 edge histogram defined by the standard [7], the quantizing of HSV and Wavelet histograms has to be defined. In HSV histogram, we set the uniform quantizing parameters in order to limit the feature size to approximately 100 bins and to privilege the finest encoding of Hue component. This led us to 45+32+32 bins in the feature representing concatenated normalized marginal distributions. We note that HSV histograms of frames proved to be an efficient feature for video similarity search [8]. As our problem is similar, the choice of this feature is straightforward. In wavelet histogram of each block we used the method presented in [9] for object based video retrieval adapted to our cell framework. To build the wavelet histogram only Y component of YUV original video color space was used. The type of wavelet decomposition is the Daubechies 9/7 2-level pyramids used in JPEG2000 standard. In [9] we computed the wavelet coefficients for at least 4 levels of decomposition. Taking into account the resolution of source video in INS we did the wavelet decomposition only in 2 levels. The feature is a weighted concatenation of two wavelet histograms per cell: $H_w(c) = (\alpha H_{LF}, (1-\alpha)H_{HF})^T$. Here H_{LF} is a histogram of LL sub-band in wavelet decomposition quantized in 2^8 bins. H_{HF} is a histogram of a mean energy in L1 norm of high frequency coefficients in wavelet decomposition $e_{HF} = 1/3(|LH(x,y)| + |HL(x,y)| + |HH(x,y)|)$. The histogram H_{HF} is also quantized in 2^8 bins. In the present experiment, $\alpha = 0.7$ according to our experience reported in [9].

2.2.2 Computation of visual dictionary

The general framework of Bag OF Words (BOW) approach in text or derived Bag Of Features (BOF) approach in image retrieval suppose the availability of dictionary or codebook. The approach largely used is the unsupervised clustering (K-means) with a large number of clusters. Here, the choice of distance between features is important both from the geometrical properties of inherent feature space and computational efficiency. In this work, we used the L1 distance for its computational efficiency and worked with dictionaries of size 1000 for HSV and combined HSV+Edge histograms, and 100 for Wavelet histograms. The latter choice was conditioned by higher dimensionality of wavelet description space. An example of codebook back projected into the concept frame is given in figure 1b for HSV histogram features. The codebook has been computed using the TRECVid development set. For each three description spaces, the final BOF is uniform: it represents the frequency of appearance of K-th word from dictionary in the query frame.

2.2.3 Search of concept

The search of video shots containing a concept of interest can be expressed as a problem of query-by-example in an image database. Here the example image Q is the keyframe containing the concept. The database DB is a set of keyframes of all video shots contained in the test set. Both the Q and DB are characterized by BOFs build on chosen feature space. Hence the problem to address is the computation of similarity measure S between BOF(Q) and BOF(R)/RϵDB. Our system supports building a BOF for a bounding box embedding the instance of a concept. In this case only those code words of region cells which are included in the bounding box are used for the BOF. Another alternative consists in computing BOF on the whole set of cells in the query frame Q. In order to compare BOF(Q) and BOF(R) we proposed an asymmetric similarity measure as a sum of the absolute values of differences between the non-empty query codewords and the corresponding image codewords ($\sum_i |q_i - r_i|\delta_{q_i}$, with $\delta_{q_i} = 0$ if $q_i = 0$, and 1 otherwise). Such an asymmetric similarity measure is justified in case of a query with spatial embedding of the signature in a bounding box. In case of comparison of BOFs of the whole frame, this measure can be

justified as it well expresses the context consistency of the concept. It would obviously work bad if the same concept appears in a different cluttered contexts.

2.3 Results

The best results were obtained in the first run using HSV features. The rank is nearly in the middle of the list of participants. The concatenated feature HSV and Edge Histogram and Wavelet coefficients show lower performances. We think that the perspective of improvement is the more complete use of spatial information with the techniques of spatial embedding emerging now in the community.

2.4 Discussion

Due to time limitations and the variable number of examples for each instance, we used only one of the example images in our queries. The experiment showed that HSV features outperform the concatenation of HSV and Edge histograms or the Wavelet features. In the future, we have to study how to take advantage of several example images for a given concept. For a given instance, we may combine the descriptors of different examples or merge the results of queries for each example image. Besides, we also used the whole image for the query. We think it helped in some cases, providing a context where the instance appeared frequently. For example, Georges Bush often appeared at the White House. It also helped for very small instances, such as logos. In future work, we have to study how to take advantage of the given precise segmentation of the object, and probably keep some information relative to context.

3 Acknowledgments

This work has been carried out in the context of the IRIM (Indexation et Recherche d'Information Multimédia) of the GDR-ISIS research network from CNRS.

Experiments presented in this paper were carried out using the Grid'5000 experimental testbed, being developed under the INRIA ALADDIN development action with support from CNRS, RENATER and several Universities as well as other funding bodies (see https://www.grid5000.fr).

References

[1] Stéphane Ayache and Georges Quénot, Video Corpus Annotation using Active Learning, In 30th European Conference on Information Retrieval (ECIR'08), Glasgow, Scotland, 30th March - 3rd April, 2008.

[2] P.H. Gosselin, M. Cord, Sylvie Philipp-Foliguet. Combining visual dictionary, kernel-based similarity and learning strategy for image category retrieval. In Computer Vision and Image Understanding, Special Issue on Similarity Matching in Computer Vision and Multimedia. Volume 110, Issue 3, Pages 403-41, 2008.

[3] D. Gorisse, M. Cord, F. Precioso, SALSAS: Sublinear active learning strategy with approximate k-NN search, Pattern Recognition, In Press, Corrected Proof, Available online 21 December 2010.

[4] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek. A comparison of color features for visual concept classification. In *ACM International Conference on Image and Video Retrieval*, pages 141–150, 2008.

[5] J. Sivic and A. Zisserman. Video google: a text retrieval approach to object matching in videos. In *ICCV'03*, volume 2, pages 1470–1477, 2003.

[6] S. Ayache, G. Quénot, J. Gensel, and S. Satoh. Using topic concepts for semantic video shots classification. In Springer, editor, *CIVR – International Conference on Image and Video Retrieval*, 2006.

[7] Chee Sun Won, Dong Kwon Park, and Soo-Jun Park. Efficient use of MPEG-7 Edge Histogram Descriptor. *ETRI Journal*, 24(1):23–30, 2002.

[8] Émilie Dumont and Bernard Merialdo. Rushes video summarization and evaluation. *Multimedia Tools and Applications, Springer, Vol.48, N1, May 2010*, 2010.

[9] Cl. Morand, J. Benois-Pineau, J.-Ph. Domenger, J. Zepeda, E. Kijak, and Ch. Guillemot. Scalable object-based video retrieval in hd video databases. In *Signal Processing: Image Communication*, volume 25, pages 450–465, July 2010.